Xmatch's Algorithm

A quick summary of the algorithm, implementation, and analysis

Arnold Rots 2025-02-27

Xmatch is based on the algorithm developed by Budavári & Szalay (2008) and Heinis, Budavári, & Szalay (2009) that was also used in the first CSC release for the CSC-SDSS cross-match, which, in turn, was used for the absolute astrometric error determination for Release 1 by Rots & Budavári (2011). See also Budavári & Loredo (2015). It differs from the match performed for Release 1 in five respects:

1. Elliptical errors were used, rather than circular approximations
2. The matches were performed on a per-Field basis, providing a better approximation to local source density. A Field is defined as the union of the fields of view of a set of spatially connected observations.
3. The probability threshold is calculated on the basis of Budavári & Szalay's (2008) self-consistency argument
4. To ameliorate issues of non-Gaussian error distribution, uncertainties surrounding the error estimates and pollution by nearby sources, the crossmatches were run twice: once with Bayes Factors based on the error ellipses (Run A) and once based on the maximum of the Bayes Factor calculated from the error ellipses and from the raw-size ellipses (Run B); raw-size ellipses are the 1-sigma PSF ellipses for point sources and the measured raw size ellipses for (compact) extended sources
5. Ambiguous matches are explicitly identified and analyzed

The following sections define the calculation of Bayes Factors, Probabilities, Probability Threshold, Ambiguity, and Match Grade; and References. I have a more complete formulation of Bayes Factors and Probabilities for matching more than two catalogs simultaneously, but that would only complicate matters.

Bayes Factors

Each source i from source set $\mathcal{L}$ is matched against each source j from source set $\mathcal{M}$. and the Bayes factor B_{ij} for each of those matches is calculated as:

$$B_{ij} = \frac{2}{\sigma_i^2(j) + \sigma_j^2(i)} \cdot \exp\left(-\frac{\psi_{ij}^2}{2\left(\sigma_i^2(j) + \sigma_j^2(i)\right)}\right)$$

If $\overline{\boldsymbol{\psi}_{ij}}$ is the vector between sources i and j, let ψ_{ij} be the angular distance in radians and ϕ_{ij} the vector's position angle. $\sigma_i(j)$ is the (Gaussian) standard deviation of the position error of source i along

the vector $\overline{\boldsymbol{\psi}_{ij}}$, i.e., the distance, in radians, along that vector from the position of i to where it intersects with its error ellipse, scaled to Gaussian standard deviation. This means:

$$\sigma_i^2(j) = \frac{a_i^2 b_i^2}{a_i^2 \sin^2(\phi_i - \theta_{ij}) + b_i^2 \cos^2(\phi_i - \theta_{ij})}$$

where a_i, a_j, b_i, and b_j are the semi axes and ϕ_i and ϕ_j the position angles of the ellipses for sources *i* and *j*, respectively. CSC2 errors (95%) are reduced by a factor 0.4085 to convert to 1-sigma.

Probabilities

We start out with the prior $P_0(0)$:

$$P_0(0) = \frac{\min(N_L, N_M)}{N_L \cdot N_M}$$

The numbers of sources in the catalogs (N_L, N_M), as well as the maximum expected number of matches, are to be scaled to the whole sky.

Next, we iterate (*k*).
Calculate for each pair (i, j) its posterior match probability p_{ij}:

$$p_{ij}(k) = \left(1 + \frac{1 - P_0(k)}{B_{ij} \cdot P_0(k)}\right)^{-1}$$

Now update P_0:

$$P_0(k+1) = \frac{\sum_{i=1}^{N_L} \sum_{j=1}^{N_M} p_{ij}(k)}{N_L \cdot N_M}$$

And iterate until:

$$\frac{P_0(k+1) - P_0(k)}{P_0(k+1)} < 10^{-3}$$

Again, the numbers of sources in the catalogs (N_L, N_M), as well as the expected number of matches, are to be scaled to the whole sky. It is prudent to limit the maximum number of iterations to something like 20: if all Bayes Factors are very small this will not converge and no harm will be done if the iterations are terminated.

Probability Thresholds

Budavári & Szalay (2008), in Section 5.3, propose a self-consistent mechanism for determining the threshold that match probabilities have to meet in order to be considered accepted. On that basis we have adopted the following considerations and criterion for acceptance.

For a given set of n pairs the iteration on the probabilities for the individual pairs is derived from the Bayes Factors and a prior that involves the sum of probabilities and the source densities. The issue here

is that the source densities introduce a scaling of the probabilities as derived from the BFs (that's why the P versus $\log(BF)$ curves are never identical) and that for assigning matches we want to apply a uniform thresholding criterion. Here is the recipe:

Assume we have a list of n source pairs with probabilities $p[i]$ and a sum $S_P = \sum_{i=1}^{n} p[i]$.
Note that we count array elements as 1-relative for clarity.

1. If $S_P < 0.2$ reject all matches; else:
2. Sort the list according to decreasing $p[i]$
3. Set $k = S_P$ (truncate)
4. Set the threshold for these n pairs to $P = \max\,(\, s \cdot p[k], 0.4\,)$
5. Accept all pairs in the list with $p[i] > P$

s is set to $s = 0.90$, to make sure no matches are missed; the absolute minimum is set at 0.4.
Experiments confirmed that these are sensible values.
Normalized separation $\frac{\psi}{\sigma}$ is used as a secondary criterion.

Ambiguity

All pairs with probability above the threshold, where neither member is a member of another pair with a valid probability, are accepted as unique (unambiguous) matches. Ambiguous matches (sources that appear in more than one accepted pair) are identified.

All ambiguously matched sources are closely inspected. In those cases where one particular match has a significantly greater probability than all others (and the same applies to the other member of that pair), that match is also accepted as unique. If the difference is still significant, but less pronounced, the match is accepted as potentially contaminated.

Match Classification

We assign six types of matches:

- Definite (d): Run A unique match with $\frac{\psi}{\sigma_A} \leq 1.7$ (i.e., within the 76% confidence region)
- Likely (l): Run A unique match with $1.7 < \frac{\psi}{\sigma_A} < 3.4$ or: no run A match but $\frac{\psi}{\sigma_B} < 3.4$ based on error ellipse (i.e., within the 99.7% confidence region)
- Definite, but potentially contaminated (c)
- Likely, but potentially contaminated (k)
- Raw (r): No Run A match, but $\frac{\psi}{\sigma_B} < 1.7$ based on raw size ellipse
- Ambiguous (a)

References

Budavári, T., & Loredo, T. J. 2015, Ann. Rev. Stat. Appl. 2015.2, 113
Budavári, T., & Szalay, A. 2008, ApJ 679, 301
Heinis, S., Budavári, T., & Szalay, A. 2009, ApJ 705, 739
Rots, A. H., & Budavári, T. 2011, ApJS 192, 8